Boosted W-tagging techniques in CMS

Andreas Hinzmann
(CERN)

for the CMS collaboration

“Using jet substructure” workshop at University of Oregon
(23-26 April 2013)
CMS searches with boosted W-tagging
Techniques for W-tagging
Validation and systematic uncertainties
Ideas for improvements
New W-tagging techniques
Discussion topics
Searches using W-tagging


  - Corresponding 8 TeV 20/fb analysis in Justin’s talk!
W/Z-tagged dijet search

- Benchmarks for model independent search:
  - \( G_{RS} \rightarrow WW/ZZ \)
  - \( W^* \rightarrow WZ \)
  - \( q^* \rightarrow qW/qZ \)

- QCD background reduced by
  - Angular distribution of dijets:
    restrict \(|\eta_1 - \eta_2| < 1.3\)
  - W/Z-tagging of jets

- Background estimated from smooth fit (S+B) to data (no need for BG MC)
- Search for \( qW/qZ \) resonances in single-W/Z-tagged sample
- Search for \( WW/WZ/ZZ \) resonances in double-W/Z-tagged sample
Hadronic $t\bar{t}$ search

- BSM models with enhanced $t\bar{t}$ production
  - Technicolor
  - Topcolor
  - Randall-Sundrum, ADD

- Benchmarks for model independent search
  - Narrow-width (1%) $Z'$ (smaller than experimental resolution in $m_{t\bar{t}}$)
  - Intermediate-width (10%) $Z'$
  - Intermediate-width (~20%) KK-gluon
  - Enhancement of the spectrum (no bump in $m_{t\bar{t}}$)

- QCD background reduced by W/top-tagging
- QCD background estimated from sideband region
- Search for bump or enhancement in $t\bar{t}$ invariant mass $m_{t\bar{t}}$
Particle Flow Jets clustered from identified particles reconstructed using all detector components.

Default jet clustering algorithms for p+p collisions:
Anti-$K_T$ with $R=0.5$ (and 0.7)
Techniques for W-tagging used

- Appy jet pruning
  - Ellis, Vermillion, Walsh (arXiv:0903.5081)
  - Improves mass resolution by removing soft, large angle particles
- Recluster each jet with Cambridge Aachen (CA) with R=0.8, requiring that each recombination satisfy the following:
  \[
  \frac{\min(p_{T1},p_{T2})}{p_{Tp}} > 0.1 \quad \Delta R_{12} < 0.5 \times \frac{m_{\text{jet}}}{p_T}
  \]
- Tag Ws (PAS JME-10-013) by
  - Butterworth et al. (arXiv:0802.2470)
  - Pruned jet mass:
    - 60 < m_{\text{jet}} < 100 GeV (hadronic ttbar)
    - 70 < m_{\text{jet}} < 100 GeV (W/Z-tagged dijets)
  - Undo last step of jet clustering to find 2 subjets
  - Mass drop:
    \[
    \mu \equiv \frac{m_{\text{leading subjet}}}{m_{\text{jet}}} < 0.4 \text{ (hadronic ttbar)}
    \]
    \[
    < 0.25 \text{ (W/Z-tagged dijets)}
    \]
Efficiency validation in ttbar

- Look for merged Ws in low mass semileptonic $t\bar{t}$ events
- Use W peak from W-tagged jets to determine
  - substructure energy correction for MC = $1.02 \pm 0.01$
  - W-tagging efficiency correction for MC = $0.97 \pm 0.03$
- Madgraph+Pythia Z2 works well with 7 TeV data
p_{T}-dependence of efficiency

- Check $p_{T}$-dependence of W-tagging efficiency by checking also $p_{T}(W+b)>400$ GeV events
- Agreement within limited statistics
- Can't check at even higher $p_{T}$ due to merging of W and b into top
- Candle for the future: semileptonic WW sample (but not at 8 TeV yet)
Validation of mistag rate

- Jet substructure in QCD background well modeled by Herwig++ MC
  - Less well in Pythia6 MC (with $p_T$-ordered shower)

- No direct impact on analysis, since QCD background modeled using data
  - However useful to understand MC tuning
**p_T-dependence of mistag rate**

- QCD background suppressed by up to a factor 150 w.r.t. signal in double-tagged sample

- p_T-dependence of mistag probability in QCD events well described by MC
  - Giving confidence that we can use MC to extrapolate also tagging efficiency to high p_T

signal efficiency ~20-50%

<table>
<thead>
<tr>
<th>Dijet Mass (GeV)</th>
<th>N (1-tag) / N (all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMS (5.0 fb⁻¹)</td>
<td>data, QCD Herwig++, q* → qZ, QCD Pythia6</td>
</tr>
</tbody>
</table>

signal efficiency ~10-30%

<table>
<thead>
<tr>
<th>Dijet Mass (GeV)</th>
<th>N (2-tag) / N (all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMS (5.0 fb⁻¹)</td>
<td>data, QCD Herwig++, W → WZ, Z → RS, G_{RS} → ZZ, G_{RS} → WW</td>
</tr>
</tbody>
</table>
Ideas for improvements

- For > 500 GeV Ws, need to use MC to extrapolate the efficiency from the region < 500 GeV, where we validate it in ttbar
  - Currently have large systematic based on Pythia6 Z2 vs. Herwig++ 23 difference
  - Can be reduced by improving shower tuning in MC
    - Aim to have single generator for showering (e.g. Pythia8) for all signals and backgrounds
    - Relies on well predictable (smooth) $p_T$-dependence of the W-tagging efficiency
  - Need to tune grooming technique to not break down when detector resolution for substructure degrades at high $p_T$
New techniques: $\tau_2/\tau_1$ and Q-jets

- $W$+jets events with jet $p_T > 200$ GeV
- Compare data to Madgraph+Pythia6 ($p_T$-ordered shower)
  - Jet substructure not so well modeled with this shower+tune
- Compare $W$+jets to $H(600 \text{ GeV}) \rightarrow WW$
  - $N$-subjettiness $\tau_2/\tau_1$ and Q-jets volatility allow better discrimination power for $W$-tagging

---

**One-pass kT axes optimization and $\beta = 1$**

**Using jet substructure**

**Boosted $W$-tagging techniques in CMS**

Andreas Hinzmann
New techniques: $\tau_2/\tau_1$

- Dijets events with jet $p_T > 700$ GeV
- Compare data to Madgraph+Pythia6 ($p_T$-ordered shower)
  - Jet substructure not so well modeled with this shower+tune
- Compare dijets to $G_{RS}(1.5$ TeV) $\rightarrow$ WW
  - N-subjettiness allows better discrimination power for W-tagging even after pruned jet mass cut
New techniques: Subjet kinematics

- More than just bump hunting...
  - Example [arXiv:1010.0676], spin and CP from ZZ→2l2q with substructure

- Example, $W_L$ identification
  - WW scattering, fitting for the $W_L W_L$ component

- What is the $\cos \theta$ resolution for subjets?

---

**Diagram:**

- **CMS simulation, $\sqrt{s} = 8$ TeV**
- **Signal W-jet, $G_{\text{RS}} \rightarrow WW$, Pythia6**
- **$m = 1$ TeV**
- **$m = 1.5$ TeV**
- **$m = 2$ TeV**
- **$m = 3$ TeV**

---

**Figure:**

- **Jet $p_T$ resolution**
  - CMS
  - $\sqrt{s} = 7$ TeV, $L = 35.9$ pb$^{-1}$
  - PFJets total systematic uncertainty
  - MG truth (c-term added)
  - $0 < |h| \leq 0.5$

- **Subjet $\Delta R$ resolution**
  - Momentum (GeV)
  - ΔR Absolute Resolution (rad)
Discussion topics

• Systematic uncertainty on tagging efficiency at high $p_T$
  • Can improve with better understanding/tuning of MC generators

• Next generation tagger
  • N-subjettiness good stand-alone variable (Q-jets also)
  • Is one variable the best for all W, Z, H?
  • Can we make use of polarization?

• How to deal with high $p_T$ and high PU at the same time?
  • Let’s discuss after Nhan’s talk
Jet performance

- Jet energy scale uncertainty: 1-2% for $p_T > 150$ GeV
- Jet calibration vs. $\eta$ better than 1% per unit of pseudorapidity for $|\eta| < 2.5$
- Jet energy resolution: 10% @ $p_T = 100$ GeV
- Jet position resolution in $\phi$ and $\eta$: ~0.01 @ $p_T = 100$ GeV
- Jet trigger efficiency: >99.9% above $p_T$ threshold
Jet substructure reconstruction

- Single particle response for $\pi^0$, $\pi^\pm$ calibrated to % precision
- Therefore jet energy corrections are only of % level
- Charge particles which do not come from the primary vertex do not enter jet reconstruction
- Jet-by-jet correction for remaining pileup contribution based on jet area and event rho (like in M. Cacciari, G. Salam PLB659(2008)119)
- When jet substructure variables are obtained (e.g. mass drop), no extra correction on the single particles entering the jet clustering algorithms are applied
  - Instead, at analysis level data/MC scale factors from ttbar are applied

- Ideas to improve jet substructure calibration in the future
  - Improve single particle response modeling in MC by an additional particle-flow cluster calibration
  - Pileup-correction for jet shape variables accounting for hadron masses (like in Soyez et al. arXiv:1211.2811)
  - Reducing or correcting for pileup can help reducing data/MC disagreements if impact of pileup in jet substructure not well modeled
Jet algorithms for reconstruction

- Full set of jet energy corrections for AK 0.5, AK 0.7
  - Based on AK due to its cone shape
- Rho computation for PU-correction based on $k_T 0.6$
  - Mainly historical
- Substruture taggers based on CA 0.8
  - CA because AK doesn’t give hard subjets when unclustering the last step
  - 0.8 because with this radius the ttbar merges to a large fraction into Ws
    - Larger radius would mainly yield merged tops
    - Smaller radius would yield merged Ws but only little statistics
- Hadronic ttbar search
  - Both ttbar spectrum reconstruction and tagging based on pruned CA 0.8
  - Small difference between AK 0.5 and CA 0.8 calibration taken into account in systematic uncertainties
- W/Z-tagged dijet search
  - Dijet spectrum reconstruction based on AK 0.5 possible due to high boost of Ws in this search
  - Tagging based on CA 0.8 because validation done in ttbar
Background estimation

- Rely on data-driven techniques for QCD background estimation
  - No need for QCD MC modeling of jet substructure

- W/Z-tagged dijet search:
  - Background can be fitted by a smooth function (inspired by PDF fitters)
    \[
    \frac{d\sigma}{dm_{jj}} = \frac{P_0(1-x)^P_1}{x^{P_2+P_3\ln(x)}}
    \]
  - Fit signal+background at the same time in the statistical analysis
  - Estimate background uncertainty from variations of the fit function

- Hadronic ttbar search:
  - Background estimated by applying the top-tagging mistag probability to signal-depleted region
  - Signal-depleted region defined by requiring only one of the two candidates to be W/top-tagged
  - Adjust the jet mass distribution to those expected from QCD MC
  - Validate this procedure using MC
• W/Z-tagged dijet search:
  • $G_{RS}$ signal MC simulated with Herwig++ (main) and Pythia6 (for checks)
  • $q^*$ and $W'$ MC simulated with Pythia6

• Hadronic ttbar search:
  • $Z'$ MC simulated with Madgraph+Pythia6
  • KK-Gluon MC simulated with Pythia8

• Observe order of $\sim$10% difference in W/Z-tagging efficiency modeling in Pythia6 and Herwig++ due to the different showering+hadronization algorithms
  • Currently taken into account as a systematic uncertainty
  • In the future would need to tune/correct the jet substructure description in MC to reduce such uncertainty
W/Z-tagged search results

- Set resonance mass limits on $q^* \rightarrow qW/qZ$
- Set cross section limits on $G_{RS} \rightarrow WWZZ$
- Translate into limit on coupling $k/M_{Pl}$ vs. resonance mass
- At high resonance masses $\sim 1.5$ TeV, the fully hadronic channel is more sensitive than semileptonic $G_{RS} \rightarrow ZZ$ resonance searches
Hadronic ttbar search results

- Set cross section limits on resonances of various width
- Fully hadronic channel sets most stringent limits are high resonance masses \( \sim 1.3 \text{ TeV} \)

---

** CMS Preliminary, 4.6 fb\(^{-1}\) at \( \sqrt{s} = 7 \text{ TeV} \)

---

** Boosted W-tagging techniques in CMS

Andreas Hinzmann

---

23 April 2013
Hadronic ttbar search results

• Look also for non-resonant change in $m_{t\bar{t}}$ spectrum

\[ S = \frac{\int_1^{+\infty} \left( \frac{d\sigma_{SM+NP}}{dm_{t\bar{t}}} \right) dm_{t\bar{t}}}{\int_1^{+\infty} \left( \frac{d\sigma_{SM}}{dm_{t\bar{t}}} \right) dm_{t\bar{t}}} \]

• Set limit on $m_{t\bar{t}} > 1 TeV/c^2$ and correct for smearing in $m_{t\bar{t}}$

<table>
<thead>
<tr>
<th></th>
<th>1+1</th>
<th>1+2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected SM tt events</td>
<td>$194 \pm 106$</td>
<td>$129 \pm 80$</td>
</tr>
<tr>
<td>Expected non-top multijet events</td>
<td>$1546 \pm 45$</td>
<td>$2271 \pm 130$</td>
</tr>
<tr>
<td>Total expected events</td>
<td>$1740 \pm 115$</td>
<td>$2400 \pm 153$</td>
</tr>
<tr>
<td>Observed events</td>
<td>$1738$</td>
<td>$2423$</td>
</tr>
<tr>
<td>tt efficiency</td>
<td>$(2.5 \pm 1.3) \times 10^{-4}$</td>
<td>$(1.6 \pm 1.0) \times 10^{-4}$</td>
</tr>
</tbody>
</table>

• Counting experiment gives $CL_S$ 95% C.L. limit: $S < 2.6$